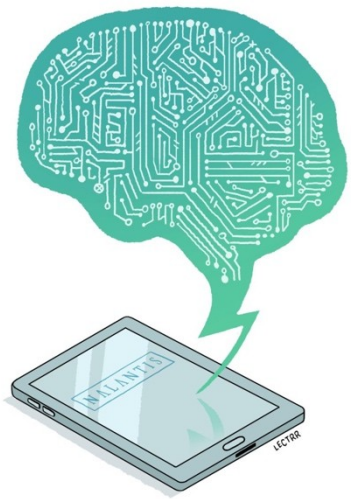
Wat is taaltechnologie?

Sinds de komst van ChatGPT eind 2022 is de populariteit van taaltechnologie spectaculair toegenomen. De media berichten onophoudelijk over elke nieuwe doorbraak en ChatGPT bereikte in een mum van tijd honderden miljoenen gebruikers. Ook jij hebt ongetwijfeld al een **chatbot** als ChatGPT gebruikt om te brainstormen, teksten te schrijven of misschien zelfs vraagstukken op te lossen.



Taaltechnologie gaat echter over meer dan chatbots zoals ChatGPT. Het is een fascinerend vakgebied dat zich bezighoudt met het ontwikkelen van computersystemen die menselijke taal kunnen begrijpen, verwerken en produceren. Het is als het ware de technologie die computers 'leert' met taal om te gaan. Denk bijvoorbeeld aan automatische vertaling of spellingcheckers in tekstverwerkers.

Taaltechnologie maakt ons leven op vele manieren gemakkelijker. Automatische vertaling helpt bijvoorbeeld mensen die een andere taal spreken elkaar te begrijpen. Spraaktechnologie maakt computers toegankelijker voor mensen die moeite hebben met typen of lezen. Ze vormt de basis voor zoekmachines, waardoor we op het internet de informatie kunnen vinden die we zoeken.

Bovendien helpt ze bij het analyseren van grote hoeveelheden tekst, wat nuttig is in onderzoek en het bedrijfsleven.

Opdracht 1

A: 1. Wat is een chatbot?

A. **Een computerprogramma dat gesprekken met mensen kan voeren**
B. Een app om spraak om te zetten in tekst
C. Een online chatroom
D. Een robot die kan praten
E. Bevestigen

B: 2. Welke van de volgende toepassingen maakt waarschijnlijk GEEN gebruik van taaltechnologie?

A. Automatische ondertiteling
B. Software voor handschriftherkenning
C. Zoekmachines op het internet
D. Spamfilters in e-mailprogramma's
E. Bevestigen

C: 3. Hoe verhoudt taaltechnologie zich tot artificiële intelligentie?
AI is een onderdeel van taaltechnologie.

A. Taaltechnologie en AI zijn volledig gescheiden vakgebieden.
B. Taaltechnologie en AI zijn precies hetzelfde.
C. Taaltechnologie is een subveld van AI dat zich specifiek richt op taalverwerking.
D. Bevestigen

D: 4. Wat maakt taaltechnologie zo uitdagend? Duid alle juiste antwoorden aan.

A. De beperkte opslagcapaciteit van computers
B. De grote variëteit aan talen en dialecten
C. De complexiteit van menselijke taal
D. De grote hoeveelheid bestaande teksten

Hoe het begon

Hoewel de hype rond ChatGPT het soms doet lijken alsof taaltechnologie een recente uitvinding is, timmeren onderzoekers al tientallen jaren aan de weg. De wieg van de taaltechnologie staat in de jaren vijftig van de vorige eeuw.

Na de Tweede Wereldoorlog, toen de spanningen tussen Oost (communisme) en West (kapitalisme) hoog opliepen, wilde de Amerikaanse overheid namelijk erg graag Russische teksten naar het Engels kunnen vertalen. Dat inspireerde enkele wetenschappers om een computersysteem voor automatische vertaling

te bouwen. Ze hoopten dat computers veel sneller en efficiënter konden werken dan menselijke vertalers, zodat de Amerikanen beter konden begrijpen wat er allemaal in de Sovjet-Unie gebeurde.

Zeventig jaar later is zowat iedereen vertrouwd met **automatische vertaling**. De technologie, die tekst of spraak van de ene taal naar de andere vertaalt zonder menselijke tussenkomst, is een van de meest populaire toepassingen van taaltechnologie. Je hebt ongetwijfeld zelf al tools als Google Translate of DeepL gebruikt om teksten naar een andere taal om te zetten. Dat gaat snel en in veel gevallen foutloos.

Toch liep de ontwikkeling van automatische vertaling in het begin niet van een leien dakje. De eerste vertaalprogramma’s zaten dan ook erg eenvoudig in elkaar. Computers probeerden zinnen woord voor woord te vertalen, wat vaak tot grappige of onbegrijpelijke resultaten leidde. Zinnen vertalen betekent tenslotte veel meer dan gewoon woorden vervangen. Ondanks het aanvankelijke enthousiasme zorgden die systemen dan ook niet voor de verhoopte doorbraak.

Opdracht 2

Lees de Franse zinnen en hun Nederlandse vertaling.

- Stel je nu voor dat een computerprogramma de Fransen zinnen naar het Nederlands probeert te vertalen door elk Frans woord door zijn Nederlandse vertaling te vervangen.
- Bij welke zinnen gaat zo’n programma de mist in, en waarom?

A. 1: Frans: Bonjour, comment ça va?
Nederlands: Hallo, hoe gaat het?

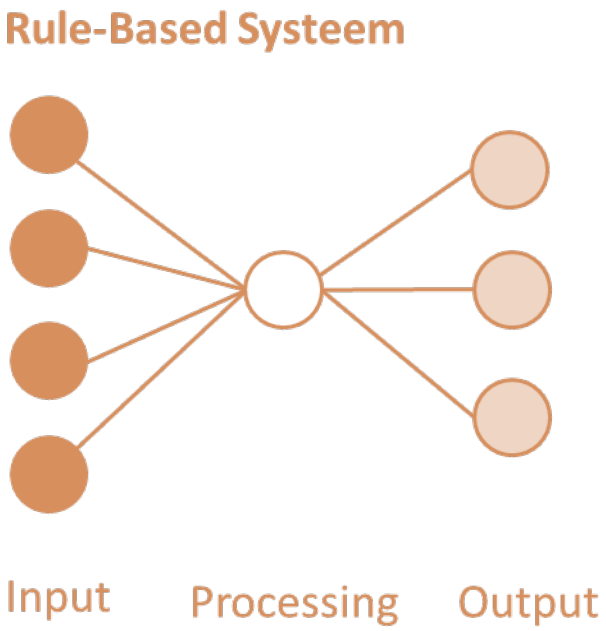
A. 2: Frans: Il a rédigé un rapport détaillé sur le projet.
Nederlands: Hij heeft een gedetailleerd verslag geschreven over het project.

A. 3: Frans: Il pleut des cordes.
Nederlands: Het regent pijpenstelen.

A. 4: Frans: Elle a deux frères et une sœur.
Nederlands: Ze heeft twee broers en een zus.
A. 5: Frans: Ils ont un bon rapport entre collègues.
Nederlands: Ze hebben een goede verstandhouding onder collega's.
A. 6 Frans: Il a acheté des cordes.
Nederlands: Hij heeft touwen gekocht.

De regels van taal

In de eerste decennia van taaltechnologie waren **regelgebaseerde systemen**, zoals hierboven, de standaard. Net zoals de eerste vertaalsoftware maakten die systemen gebruik van vooraf gedefinieerde regels en grammatica's die een taal konden analyseren en genereren. De regels werden opgesteld door taalkundigen en programmeurs. Ze bevatten gedetailleerde beschrijvingen van hoe woorden gecombineerd konden worden om zinnen te vormen, en hoe bepaalde combinaties van woorden geïnterpreteerd moesten worden.



Een bekend voorbeeld van zo'n regelgebaseerd systeem is ELIZA, een vroege chatbot uit de jaren 1960. ELIZA imiteerde een psycholoog in gesprek met een patiënt. Het programma gebruikte een set eenvoudige regels om uitspraken van zijn patiënt te interpreteren en er gepast op te antwoorden. Zo checkte het elke zin op een reeks sleutelwoorden. Als de cliënt bijvoorbeeld een zin intypte met de woorden *alike* of *same*, reageerde ELIZA automatisch met '*In what way?*'. Als ELIZA niet meteen wist wat aan te vangen met een bepaalde uitspraak van de patiënt, antwoordde ze met een nietszeggende zin, zoals '*I see*' of '*Please go on*'. Zoals het voorbeeld hieronder toont, leidde dat niet meteen tot boeiende gesprekken.

Patiënt: *Men are all alike.*
ELIZA: *In what way?*
Patiënt: *They're always bugging us about something or other.*
ELIZA: *Can you think of a specific example?*
Patiënt: *Well, my boyfriend made me come here.*
ELIZA: *Your boyfriend made you come here.*
Patiënt: *He says I'm depressed much of the time.*
ELIZA: *I am sorry to hear you are depressed.*

Regelgebaseerde systemen hadden af te rekenen met verschillende uitdagingen. Omdat taal complex is en voortdurend verandert, is het zo goed als onmogelijk om alle mogelijke regels in software te vatten. Dat had vaak

onhandige vertalingen of gesprekken als resultaat. Zo hebben regelgebaseerde systemen moeite met woorden die meerdere betekenissen kunnen hebben. Die uitdagingen zorgden er uiteindelijk voor dat aan het einde van de vorige eeuw regelgebaseerde software in de taaltechnologie plaats moest ruimen voor meer geavanceerde benaderingen.

Sentimentanalyse

Naast chatbots en automatische vertaling is ook **sentimentanalyse** een populaire toepassing van taaltechnologie. Daarbij proberen computers te begrijpen of een tekst positief, negatief of neutraal is. Als je bijvoorbeeld een boze tweet schrijft over huiswerk, kan sentimentanalyse automatisch herkennen dat je bericht negatief is.

Sentimentanalyse is bijzonder populair bij bedrijven. Zij gebruiken de technologie om te begrijpen hoe mensen denken over hun producten of diensten. Ze verzamelen bijvoorbeeld alle berichten op sociale media die hun merknaam vermelden, en gebruiken sentimentanalyse om te bepalen welke opmerkingen positief en negatief zijn. Zo kunnen ze de populariteit van hun merk door de tijd heen in het oog houden en de opmerkingen gebruiken om hun producten verder te verbeteren.

Opdracht 3

Laten we een regelgebaseerd systeem voor sentimentanalyse onder de loep nemen. Het systeem werkt met een klein woordenboek, dat ons vertelt of woorden een positieve of negatieve betekenis hebben. Het combineert dat woordenboek met enkele eenvoudige regels om de betekenis van een volledige zin te achterhalen.
Veronderstel dat Apple dit eenvoudige systeem gebruikt om de meningen over de iPhone te monitoren. Het heeft daartoe een groot aantal tweets verzameld, waaronder de tien voorbeelden hieronder.

Woordenboek

- Positieve woorden: *goed, geweldig, fantastisch, prachtig, uitstekend*
- Negatieve woorden: *slecht, vreselijk, verschrikkelijk, vervelend, teleurstellend*

Regels

- Ontkenning: als een ontkenning (*niet* of *geen*) voor een positief woord staat, wordt het sentiment negatief en vice versa.
-
- Versterking: woorden als *zeer* of *extreem* versterken de polariteit van het volgende woord.
- Combinatie: de sentimenten van meerdere woorden worden opgeteld om de algemene toon te bepalen. Twee

positieve woorden maken een tweet bijvoorbeeld heel positief.

Beantwoord voor elke tweet de volgende vragen.

- Analyseert Apples software voor sentimentanalyse de tweet als heel negatief, negatief, neutraal, positief of heel positief?
- Is die analyse correct?
- Als de analyse niet correct is, waarom gaat de software in de fout?

A. 1: De nieuwe iPhone is echt geweldig!

Neutraal
Negatief
Heel negatief
Heel positief
Positief
Bevestigen

A. 2: Ongelooflijk hoe vreselijk snel die nieuwe iPhone is!

Positief
Negatief
Neutraal
Heel negatief
Heel positief
Bevestigen

A. 3: De nieuwe iPhone: fantastisch scherm, uitstekende batterijduur.

Positief
Negatief
Heel negatief
Heel positief
Neutraal
Bevestigen

A. 4: Nu is het scherm van mijn iPhone alweer kapot. Geweldig!

Positief
Heel negatief
Heel positief
Negatief
Neutraal

Bevestigen

A. 5: Ik had verwacht dat de nieuwe iPhone fantastisch zou zijn, maar hij is gewoon goed.

Heel negatief
Neutraal
Negatief
Heel positief
Positief
Bevestigen

A. 6: Wil er iemand mijn oude iPhone kopen?

Positief
Neutraal
Negatief
Heel positief
Heel negatief
Bevestigen

A. 7: Ik heb nog geen slechte ervaring gehad met mijn nieuwe iPhone.

Heel positief
Heel negatief
Negatief
Neutraal
Positief
Bevestigen

A. 8: De batterijduur van de nieuwe iPhone is echt teleurstellend.

Negatief
Heel positief
Positief
Neutraal
Heel negatief
Bevestigen

A. 9: Is het niet vreselijk hoe duur de nieuwste iPhone is?

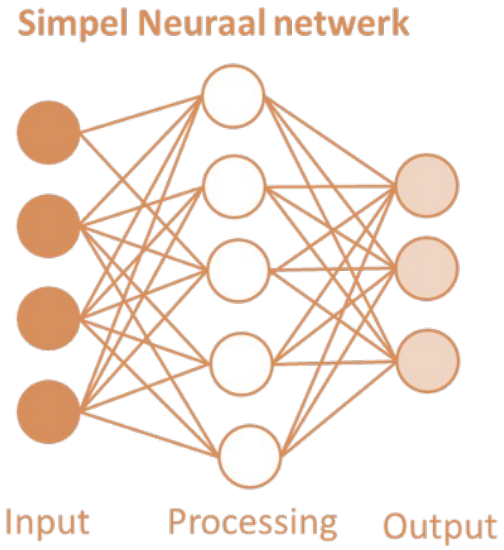
Positief
Heel positief
Heel negatief
Neutraal
Negatief
Bevestigen

A. 10: Die nieuwe iPhone! □□□

Positief
Heel negatief
Heel positief
Neutraal
Negatief

<u>Machines die leren</u>

Het is nu wel duidelijk dat systemen die werken op basis van regels er nooit in zullen slagen de complexiteit van taal te vatten. Bovendien vergt het enorm veel tijd om zulke lijsten van regels aan te leggen en bij te houden. Eind jaren tachtig van de vorige eeuw begon dan ook een nieuwe benadering van taaltechnologie furore te maken: ***machine learning***.



Bij *machine learning*, of automatisch leren, is het niet nodig dat de ontwikkelaars van taaltechnologische systemen zelf alle relevante regels in de software programmeren. In plaats daarvan kan de computer zelf leren uit grote hoeveelheden teksten. Dat zijn zogenaamde trainingsdata. Om bijvoorbeeld een systeem voor automatische vertaling te bouwen, laten we de computer als trainingsdata miljoenen voorbeelden van zinnen en hun correcte vertaling analyseren.

Om een systeem voor sentimentanalyse te ontwikkelen, geven we de computer een reeks voorbeelden van zinnen en hun correcte sentiment. Op die manier kan de computer zelf patronen ontdekken. Zo kan hij afleiden hoe elk woord vertaald wordt of welke emotionele lading een woord heeft, en welke invloed de context daarbij speelt.

Die aanpak is veel efficiënter en flexibeler dan de eerdere, regelgebaseerde benadering. Systemen die werken op basis van *machine learning* gaan veel flexibeler om met de natuurlijke variatie in taal. Ze kunnen zelfs nieuwe woorden of uitdrukkingen leren zonder dat een mens elke keer nieuwe regels hoeft toe te voegen. Het is een beetje zoals een kind taal leert: niet door alle grammaticaregels uit het hoofd te leren, maar door veel voorbeelden te horen en te lezen en daarin patronen te herkennen.

Opdracht 4

Leer vertalen uit het Zorbisch, een onbestaande taal.

- Leer de taal kennen zoals moderne systemen voor machinevertaling dat doen: zonder een grammatica of woordenboek, maar enkel op basis van voorbeeldzinnen en hun vertaling. Zo zul je zien dat je computers helemaal geen woordenboek of grammaticale regels nodig hebben om correct te leren vertalen.
- Lees de voorbeeldzinnen.

Miki bula	Ik eet.
Taso bula	Jij eet.
Miki dorna	Ik slaap.
Loki miki	Ik ben blij.
Loki taso	Jij bent blij.
Zuma bula kafi	De hond eet voer.
Zuma dorna hapi	De hond slaapt rustig.

Probeer nu op basis van de voorbeeldzinnen te achterhalen wat de volgende Zorbische woorden betekenen.

A. 1: Miki
Juiste antwoord
Ik
A. 2: Bula
Juiste antwoord
eet
A. 3: Loki
Juiste antwoord
blij zijn

A. 4: Taso
Juiste antwoord
jij

A. 5: Dorna
Juiste antwoord
slaap(t)

A. 6: Zuma
Juiste antwoord
de hond

Backtranslation

Het is dus mogelijk enkel op basis van voorbeelden nieuwe zinnen te vertalen naar een taal die je nooit eerder hebt gezien. Op een gelijkaardige manier werkt *machine learning*. De computer zoekt naar patronen in grote hoeveelheden voorbeelden en past die kennis toe op nieuwe zinnen.

Natuurlijk heb je in de vorige oefening slechts een klein stukje Zorbisch geleerd. Om alle mogelijke zinnen accuraat naar een nieuwe taal te vertalen, heeft een computer miljoenen voorbeelden nodig. Maar het basisprincipe blijft hetzelfde.

Ook als we ze heel veel voorbeelden hebben laten zien, vertalen computers sommige zinnen niet helemaal correct. Soms komt dat omdat ze niet genoeg voorbeelden van een bepaalde zin of constructie hebben gekregen, soms omdat verschillende talen nu eenmaal anders in elkaar zitten.

Opdracht 5

Experimenteer met *backtranslation*: vertaal Nederlandse zinnen naar een andere taal, en dan weer terug.
Gebruik daarvoor een tool voor automatische vertaling, zoals Google Translate of DeepL.

A. 1: Geef de vertaaltool enkele eenvoudige zinnen, zoals *Ik ga naar school* of *Ik woon in België.*

- Vertaal die zinnen naar een willekeurige vreemde taal, en weer terug.
- Wat merk je?

A. 2: Laat de tool nu de volgende zinnen naar een andere Europese taal vertalen, zoals het Frans of het Deens, en weer terug.

1. Nu komt de aap uit de mouw.
2. Ik heb mijn fiets voor een appel en een ei gekocht.
3. Gisteren regende het pijpenstelen.
4. Omdat hij geen Frans sprak, viel hij door de mand.

- Wat gebeurt er?
- Waarom?

A. 3: Doe nu hetzelfde, maar met talen uit andere werelddelen, bijvoorbeeld Japans of Swahili.

- Wat merk je nu?
- Hoe kun je dat verklaren?

<u>Vaktermen</u>

Automatische vertaling: software die tekst of spraak van de ene taal naar de andere kan vertalen zonder menselijke tussenkomst. Bekende voorbeelden zijn Google Translate en DeepL.

Chatbot: computerprogramma dat ontworpen is om gesprekken met mensen te voeren via tekst of spraak. Chatbots kunnen vragen beantwoorden, informatie geven of eenvoudige taken uitvoeren.

Machine learning: tak van AI waarbij de software zelf leert van data, zonder expliciet geprogrammeerd te worden voor elke taak. De software verbetert zijn prestaties naarmate hij meer data verwerkt.

Regelgebaseerd systeem: AI-programma dat een vaste set regels volgt om beslissingen te nemen. Bij taaltechnologische software gaat het bijvoorbeeld om de grammaticale regels van de relevante taal. Die regels zijn vooraf door mensen geprogrammeerd.

Sentimentanalyse: techniek waarbij AI wordt gebruikt om de emotionele toon of mening in een tekst te bepalen, bijvoorbeeld of een recensie positief of negatief is.